*Appl. Math. Lett.* Vol. 5, No. 3, pp. 3–6, 1992
Printed in Great Britain. 

0893-9659/92 $5.00 + 0.00


# CONVEX TREE REALIZATIONS OF PARTITIONS

ANDREAS DRESS
Fakultät für Mathematik, Universität Bielefeld
Postfach 8640, W-4800 Bielefeld 1, Germany
MICHAEL STEEL
Zentrum für Interdisziplinäre Forschung, Bielefeld
Wellenberg 1, 4800 Bielefeld 1, Germany

*(Received August 1991)*

**Abstract**—Given a collection $P$ of partitions of a label set $L$ a problem arising in biological and linguistic classification is deciding whether there is a "tree-structure" on $L$ which is "compatible" with $P$. While this problem is, in general, $NP$-complete we show that it has a polynomial time solution if the number of sets in each partition is at most three.

## 1. INTRODUCTION

A problem arising in certain branches of classification such as taxonomy [1] is to determine whether a collection of partitions of some label set has a tree-like representation in a sense which we will define shortly. In order to do this, and to place this question in a context which relates it to a simpler problem in which the partitions are already endowed with some tree-like structure we must first make a number of definitions.

(1) Given a tree $T = (V(T), E(T))$ and a subset $A$ of $V(T)$ let $< A >_T$ denote the minimal connected subset of $V(T)$ which contains $A$.
(2) A *semilabelled tree on* $L$ is a pair $\tau = (T, f)$ where $T$ is a tree, $f : L \to V(T)$ is a map, and if $v \notin f(L)$ then $\deg(v) > 2$. Two such pairs $\tau = (T, f)$ and $\tau' = (T', f')$ are considered identical if there is a tree isomorphism $H : V(T) \to V(T')$ with $f' = H \circ f$.
(3) A partition $X$ of $L$ into disjoint subsets $a, b, \ldots$, is *convex* on a semilabelled tree $\tau = (T, f)$ if for all $a, b \in X, a \neq b, < f(a) >_T \cap < f(b) >_T = \emptyset$.
(4) A collection $P$ of partitions of $L$ is *compatible* if there exists a semilabelled tree $\tau$ on which each partition is convex (we say $\tau$ is compatible with $P$).
(5) Given two trees $T, T'$ a function $h : V(T) \to V(T')$ is a *contraction* if $T'$ is obtained (up to isomorphism) by collapsing edges of $T$, and $h$ is the induced vertex identification map.
(6) For semilabelled trees $\tau = (T, f), \tau' = (T', f')$ we say $\tau$ is a *refinement* of $\tau'$ if there is a contraction $h : V(T) \to V(T')$ satisfying $f' = h \circ f$.
(7) A collection $S$ of semilabelled trees on $L$ are *compatible* if there exists a semilabelled tree $\tau$ which is a refinement of each semilabelled tree in $S$ (we say $\tau$ is compatible with $S$).
(8) A semilabelled tree $\tau = (T, f)$ on $L$ defines a partition $\pi(\tau)$ of $L$ by setting $\pi(\tau) = \{f^{-1}(\{v\}) : v \in V(T), f^{-1}(\{v\}) \neq \emptyset\}$.
(9) Given a semilabelled tree $\tau = (T, f)$ deleting any edge $e \in E(T)$ partitions $V(T)$ into two connected subsets; applying $f^{-1}$ to these sets partitions $L$ into at most two sets. Let $\beta_\tau$ be the set of bipartitions of $L$ which can be generated in this way from $\tau$. For a set $S$ of semilabelled trees on $L$ let $B(S) = \cup_{\tau \in S} \beta_\tau$.
(10) Following Buneman [2] a set $B = \{\sigma_1, \ldots, \sigma_k\}$ of bipartitions of $L$ defines a connected graph $G[B] = (V, E)$, and a labelling function $f : L \to V$ as follows: $V$ consists of all sets $v = \{S_1, \ldots, S_k\}$ where $S_i \in \sigma_i$ and $S_i \cap S_j \neq \emptyset$ for all $i, j$. Two such sets $v, v'$ are the ends

of an edge in $E$ precisely if $|v \cap v'| = k - 1$. For $v \in V, v = \{S_1, \ldots, S_k\}$, let $I_v = \cap_i S_i$. Then it can be checked that $\{I_v : v \in V, I_v \neq \emptyset\}$ partitions $L$, and thus the function

$$f : L \to V,$$
$$f(x) = \{S_1(x), \ldots, S_k(x)\}, \text{ with } x \in S_i(x) \in \sigma_i,$$

is well defined and its range contains all $v \in V$ of degree at most two.

In taxonomy, a partition of $L$ is called an "unordered qualitative character", while a semilabelled tree is called an "undirected cladistic character". In taxonomic and other applications an important class of semilabelled trees $(T, f)$ are *phylogenetic trees*, for which $f$ is a bijection from $L$ onto the degree one vertices of $T$. In case all the remaining vertices of $T$ have degree 3, $(T, f)$ is called a *binary* (or *nondegenerate*) *phylogenetic tree.*

## 2. PRELIMINARIES

The following result summarizes the fundamental relationships between the definitions introduced above. Most of (2) is due to Buneman [2]; see also Barthélemy [3]. Result (3a) follows from the first part of (1) and the observation that a bipartition $\sigma$ is convex on $\tau$ precisely if $\sigma \in \beta_\tau$, while (3b) follows from (2b). Result (5) was stated by Buneman [4] and Meacham [1]. The remaining results are largely part of the folklore (see, for example [5]).

THEOREM 1.

(1) *$\tau$ is a refinement of $\tau'$ precisely if $\beta_\tau \supseteq \beta_{\tau'}$. Furthermore, $\beta_\tau = \beta_{\tau'}$ implies $\tau = \tau'$; thus semilabelled trees are partially ordered by refinement.*

(2a) *Two bipartitions $\{A_1, A_2\}, \{B_1, B_2\}$ are compatible precisely if $\emptyset \in \{A_i \cap B_j : i, j \in \{1, 2\}\}$.*

(2b) *$G[B]$ is a tree precisely if $B$ is pairwise compatible. In this case $\tau(B) := (G[B], f)$ is a semilabelled tree which is compatible with $B$; $\beta_{\tau(B)} = B$; any other semilabelled tree compatible with $B$ is a refinement of $\tau(B)$, and (assuming $B$ has no repeated bipartitions) $G[B]$ has $|B| + 1 < 2|L| - 2$ vertices.*

(3a) *A collection $S$ of semilabelled trees is compatible (with $\tau$) if and only if the induced bipartitions $B(S)$ are compatible (with $\tau$).*

(3b) *A collection $B$ of bipartitions is compatible (with $\tau$) if and only if $B$ is pairwise compatible (with $\tau$).*

(4) *A collection $P$ of partitions is compatible (with $\tau$) if and only if there exists a collection $S_P$ of semilabelled trees which is compatible (with $\tau$) and such that $P = \{\pi(\tau) : \tau \in S_P\}$. Furthermore in case $P$ is compatible we may insist, for each $(T, f) \in S_P$, that $f$ is surjective.*

(5) *A collection $P$ of partitions is compatible if and only if the intersection graph of the set system $\bigcup_{X \in P} X$ can be transformed into a chordal graph by introducing additional edges, but subject to the restriction that vertices $a, a'$ remain non-adjacent if for some $X \in P, a, a' \in X$.*

(6) *If a collection of partitions or of semilabelled trees are compatible then the collection is compatible with a binary phylogenetic tree (as defined in the introduction).*

Combining 3(a) and 3(b) gives the following result, due to Estabrook, Johnson and McMorris [6].

COROLLARY 1. *"The Pairwise Compatibility Theorem" A collection $S$ of semilabelled trees is compatible if and only if it is pairwise compatible. Indeed, by Theorem 1, there is a unique semilabelled tree which is both compatible with $S$ and minimal with respect to refinement, namely $(G[B(S)], f)$.*

Applying Theorem 1(5) it is easily shown that two partitions are compatible precisely if the bipartite intersection graph of the two partitions is acyclic, a result proved explicitly by Estabrook and McMorris [7].

## 3. AN EXTENSION

By the above corollary, the compatibility of a set of semilabelled trees can be established in polynomial time. In particular, as is well known, deciding whether a collection $B$ of bipartitions of $L$ is compatible, and constructing a compatible tree, can be carried out in polynomial time (see [8] or [9] for an $0(|L| \times |B|)$ algorithm).

Polynomial time also applies if a bound is placed on the number of partitions. However, deciding whether an arbitrary set of partitions is compatible is $NP$-complete, even if each set in each partition has cardinality at most two [10]. A natural question then is whether the compatibility of sets of partitions, each having a bounded number of sets, can be decided in polynomial time. Corollary 2 (below) answers this affirmatively in case each partition has at most three sets. It is not known whether the same applies if each partition has at most $k$ sets (for fixed $k$), even when $k = 4$. Recall that a star-shaped tree is a tree having all but one of its vertices of degree 1.

THEOREM 2. *For a collection $P$ of partitions of $L$ constructing a compatible collection $S_P$ of star-shaped semilabelled trees such that $P = \{\pi(\tau) : \tau \in S_P\}$, or deciding that no such collection exists, can be achieved in $0(|L| \times p^2)$ time, where $p = |\bigcup_{X \in P} X|$.*

PROOF. Regard the sets occuring in at least one partition in $P$ as the vertices of a graph $G$, in which two sets $A, B$ are joined by an edge of $G$ precisely if $\{A, A'\}$ and $\{B, B'\}$ are incompatible partitions (here $'$ denotes complement). Note that each partition is an independent set of vertices of $G$ (that is, no two vertices are adjacent). Also, deciding whether an edge exists between two vertices takes $0(|L|)$ time, and there are $0(p^2)$ pairs of vertices, so $G$ can be constructed in $0(|L| \times p^2)$ time. For $X \in P$ there is a natural bijection $\lambda_X$ from $X^* := \{V : V \subseteq X, |V| \geq |X| - 1\}$ to $\{\beta_\tau : \tau \text{ is star-shaped}, \pi(\tau) = X\}$, namely

$$\lambda_X(\{A_1, \ldots, A_s\}) := \{\{A_1, A_1'\}, \ldots, \{A_s, A_s'\}\}, s = |X| \text{ or } |X| - 1.$$

By the definition of $G$, if $V_X \in X^*$ for all $X \in P$, then $\cup \lambda_X(V_X)$ is pairwise compatible precisely if $\cup V_X$ is an independent set of vertices in $G$. Thus by Theorem 1(3) the collection $S_P$ in the statement of the theorem exists precisely if $G$ has an independent set of vertices containing at least $|X| - 1$ vertices from each $X \in P$. We now describe a simple procedure which finds such an independent set if it exists. Essentially we describe a way of building up a set $I$, which is always an independent set, by testing the effect of adding one more new vertex. Set $I = \emptyset$ and while there exists $X \in P$, with $|I \cap X| < |X| - 1$ select $x \in X - I$, and let $A := \{x\}, B := \emptyset$. Apply the following rule:

$\mathcal{R}$ : While

(1) $A \cap B = \emptyset$ and

(2) there exists $a \in A$ which is adjacent to $y \notin B$,

then replace $A$ and $B$ by $A \cup \bigcup_{y \in Y \in P} (Y - \{y\})$ and $B \cup \{y\}$, respectively.

Eventually, in $0(p)$ steps, either (1) or (2) fails. If (2) fails but not (1), then replace $I$ by $I \cup A$. If (1) fails replace $A$ and $B$ by $\bigcup_{x \in X \in P} X - \{x\}$ and $\{x\}$ respectively, and apply rule $\mathcal{R}$ again, and in this case if (2) fails but not (1) then again replace $I$ by $I \cup A$, while if (1) fails then no independent set of the type required can exist (since it would have to simultaneously include and exclude $x$). Thus, provided this does not occur, $|I|$ is enlarged by at least one element, and, by induction, all the elements of $I$ are non-adjacent. Thus, in $0(p)$ steps, the procedure described gives the required independent set.

Since every semilabelled tree $\tau$ with $|\pi(\tau)| \leq 3$ is star shaped, combining Theorem 2 with part (4) of Theorem 1, and applying a reconstruction algorithm, like that described in [9], we obtain the following.

COROLLARY 2. *If $P$ is a collection of partitions of $L$ into at most three disjoint sets then constructing a semilabelled tree which is compatible with $P$, or deciding that no such tree exists can be achieved in $0(|L| \times |P|^2)$ time.*

## REFERENCES


1. C.A. Meacham, Theoretical and computational considerations of the compatibility of qualitative taxonomic characters, In *Numerical Taxonomy* (J. Felsenstein, Editor), NATO ASI Series Vol. G1, Springer-Verlag Berlin Heidelberg, pp. 304–314, (1983).
2. P. Buneman, The recovery of trees from measures of dissimilarity, In *Mathematics in the Archaeological and Historical Sciences* (F.R. Hodson, D.G. Kendall and P. Tautu, Editors), Edinburgh University Press, Edinburgh, pp. 387–395, (1971).
3. J.P. Barthélemy, From copair hypergraphs to median graphs with latent vertices, *Discr. Math.* **76**, 9–28 (1989).
4. P. Buneman, A characterization of rigid circuit graphs, *Discr. Math.* **9**, 205–212 (1974).
5. H.-J. Bandelt and A. Dress, Reconstructing the shape of a tree from observed dissimilarity data, *Adv. Appl. Math.* **7**, 309–343 (1986).
6. G.F. Estabrook, C.S. Johnson, Jr. and F.R. McMorris, An algebraic analysis of cladistic characters, *Discr. Math.* **16**, 141–147 (1976).
7. G.F. Estabrook and F.R. McMorris, When are two taxonomic characters compatible?, *J. Math. Biol.* **4**, 195–299 (1977).
8. D. Gusfield, Efficient algorithms for inferring evolutionary trees, *Networks* **21**, 19–28 (1991).
9. C.A. Meacham, A manual method for character compatibility analysis, *Taxon* **30**, 591–600 (1981).
10. M.A. Steel, The complexity of reconstructing trees from qualitative characters and subtrees, (submitted to *J. Classification*) (1991).